SCIENCE CHINA
Life Sciences

• RESEARCH PAPERS •

February 2011 Vol.54 No.2: 189–193
doi: 10.1007/s11427-010-4129-7

Functional modules with disease discrimination abilities for various cancers

YAO Chen[1†], ZHANG Min[2†], ZOU JinFeng[1], LI HongDong[1], WANG Dong[1], ZHU Jing[1] & GUO Zheng[1,2*]

[1]*Bioinformatics Centre, School of Life Science, University of Electronic Science and Technology of China, Chengdu 610054, China;*
[2]*College of Bioinformatics, Harbin Medical University, Harbin 150086, China*

Received February 6, 2009; accepted September 22, 2009

Selecting differentially expressed genes (DEGs) is one of the most important tasks in microarray applications for studying multi-factor diseases including cancers. However, the small samples typically used in current microarray studies may only partially reflect the widely altered gene expressions in complex diseases, which would introduce low reproducibility of gene lists selected by statistical methods. Here, by analyzing seven cancer datasets, we showed that, in each cancer, a wide range of functional modules have altered gene expressions and thus have high disease classification abilities. The results also showed that seven modules are shared across diverse cancers, suggesting hints about the common mechanisms of cancers. Therefore, instead of relying on a few individual genes whose selection is hardly reproducible in current microarray experiments, we may use functional modules as functional signatures to study core mechanisms of cancers and build robust diagnostic classifiers.

differentially expressed gene, functional module, classification, diverse cancers

Citation: Yao C, Zhang M, Zou J F, *et al*. Functional modules with disease discrimination abilities for various cancers. Sci China Life Sci, 2011, 54: 189−193, doi: 10.1007/s11427-010-4129-7

With data from microarrays, one of the most important tasks is to identify differentially expressed genes (DEGs) which would be "significant" or "important" enough for the follow-up studies [1–3]. However, DEG lists produced from different studies for a cancer are usually highly inconsistent [4,5]. As recently demonstrated by us [6], even when technical measurement variations are small, DEG lists correctly determined by a statistical method such as SAM (Significance Analysis of Microarrays) [1] from different microarray studies for a cancer still tend to be highly inconsistent. Under such situations, selecting DEGs solely based on statistical cut-off criteria (e.g., a false discovery rate (FDR) level) becomes less relevant [7–9]. However, functional modules significantly enriched with DEGs tend to be robust to the uncertainty introduced by DEG selection [10,11], as also formally demonstrated by us recently [12].

In this work, we used another approach to identify functional modules with high disease discrimination abilities for cancers. Our results demonstrated that many functional modules have high disease discrimination abilities for the five types of cancers originating in different tissues, indicating that each cancer involves the disruption of various cellular processes and genes in these functions undergo broad expression changes [13,14]. Especially, the modules shared by the five types of cancers also cover a wide range of functions, suggesting the mechanisms shared by various cancers. Obviously, understanding the mechanisms shared by various cancers has important diagnostic implications. For example, the selected modules can be used as functional signatures to build robust cancer diagnostic classifiers [15].

†Contributed equally to this work
*Corresponding author (email: guoz@ems.hrbmu.edu.cn)

1 Materials and methods

1.1 Datasets

The data analyzed in this study was downloaded from Gene Expression Omnibus database (GEO) [16], as summarized in Table 1.

Table 1 Seven cancer datasets used in this study

Data (reference)	Sample size	(Primary tumor versus normal)	Genes
Lung cancer [17]	18	(13 versus 5)	9476
Lung cancer [18]	38	(21 versus 17)	8751
Colorectal cancer [19]	23	(15 versus 8)	19772
Colorectal cancer [20]	64	(32 versus 32)	19772
Prostate cancer [21]	103	(62 versus 41)	12953
Gastric cancer [22]	132	(103 versus 29)	14123
Liver cancer [23]	156	(82 versus 74)	9350

The cDNA data was log2-transformed and then normalized to be with median 0 and standard deviation 1 per array, as adopted in Oncomine database [24]. The CloneIDs with missing rates above 20% were deleted, and the other missing values were replaced by using the *k*-Nearest Neighbor (KNN) imputation algorithm (*k*=15) [17]. The Affymetrix GeneChip data was preprocessed by RMA (Robust Multi-array Analysis). The most recent (July 2008) SOURCE database [25] was used for annotating CloneID to GeneID.

1.2 Evaluating functional expression changes of cancers by KNN classifiers

Based on each "Biological Process" term of Gene Ontology (GO) [26] (downloaded in October 2008), we built a KNN classifier [27] to classify samples in a dataset, where the Euclidean distance between two samples was calculated based on, in these samples, the expression values of the genes annotated to the GO term. Then, for each sample, we found its k (=3) nearest samples and classified this sample with the majority vote of the labels of its k nearest neighbors. The performance of a KNN classifier was evaluated by a leave-one-out cross-validation (LOOCV) procedure, in which each sample was left out in turn to calculate the accuracy rate (the percentage of the correct predictions) [28]. When a classifier achieved an accuracy rate above a threshold (e.g., 90%), we referred to the corresponding GO term as a functional module, indicating that the genes described by it can characterize the disease.

GO terms are hierarchically organized and thus are highly redundant. Some algorithms were designed to reduce this redundancy, for example, by eliminating genes mapped to significant GO terms from more general GO terms [29] or by weighting genes based on the scores of neighboring GO terms [30]. However, all these methods rely on specific assumptions which need to be further evaluated. Here, for the purpose of this study, we applied a simple method: Eliminating a functional module if one of its offspring terms can also achieve the given accuracy rate level. The specific terms identified by this method are sufficient to support our conclusion that biological pathways are widely disturbed in cancers.

1.3 Up/down-regulated genes enrichment in the modules

To test whether the shared modules were significantly up- (or down-) regulated in a cancer, we first found the up- (or down-) regulated genes in each module. In a cancer dataset, a gene was defined as up- (or down-) regulated if its mean expression value in the cancer samples is greater (or less) than that in the normal samples. Then, we calculated the fraction of up- (or down-) regulated genes in each module and used the hypergeometric distribution model to calculate its statistical significance (*P*-value).

$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}},$$

where x is the number of the up- (or down-) regulated genes in the module and N is the number of all the measured genes in the module. K is the number of up- (or down-) regulated genes and M is the total number of genes in the dataset.

2 Results

2.1 Functional modules with high disease discriminating abilities

Using KNN classifiers, for prostate, liver, lung, gastric and colorectal cancers respectively, we found 26, 60, 135, 182 and 334 functional modules with classification accuracy rates exceeding 90%. This result demonstrated that functional modules with high disease discriminating abilities cover a wide range of functions globally perturbed in cancer.

Importantly, setting the threshold of disease classification accuracy above 85%, we found seven modules shared by all the seven datasets for five types of cancers (Table 2), suggesting hints about common mechanisms of these cancer types.

2.2 Seven functional modules shared by different cancer types

Hanahan and Weinberg [31] suggested that six functional hallmarks are shared in common by perhaps all types of

Table 2 Seven functional modules with classification accuracy >85% for all the seven datasets[a)]

GO term	Prostate	Liver	Gastric	Lung [17]	Lung [18]	Colorectal [19]	Colorectal [20]
Positive regulation of I-kappaB kinase/ NF-kappaB cascade	0.88 (7.11×10^{-3})$_d$	0.93 (1.29×10^{-1})$_d$	0.98 (7.35×10^{-1})$_d$	0.94 (1.63×10^{-2})$_d$	0.87 (2.21×10^{-2})$_d$	1.00 (6.05×10^{-2})$_u$	1.00 (4.04×10^{-2})$_d$
Transcription initiation from RNA polymerase II promoter	0.91 (2.10×10^{-3})$_u$	0.86 (3.49×10^{-4})$_u$	0.94 (5.46×10^{-7})$_u$	0.89 (4.69×10^{-3})$_u$	0.97 (2.22×10^{-5})$_u$	0.97 (7.29×10^{-5})$_u$	0.91 (1.59×10^{-11})$_u$
Cell cycle arrest	0.85 (8.44×10^{-1})$_u$	0.91 (3.84×10^{-1})$_u$	0.99 (3.12×10^{-3})$_u$	0.94 (4.79×10^{-1})$_u$	1.00 (3.84×10^{-1})$_u$	0.98 (9.63×10^{-1})$_u$	0.91 (2.43×10^{-1})$_u$
Water transport	0.88 (1.36×10^{-3})$_d$	0.90 (8.05×10^{-4})$_d$	0.89 (3.07×10^{-2})$_d$	0.94 (7.40×10^{-4})$_d$	1.00 (6.21×10^{-3})$_d$	1.00 (6.77×10^{-2})$_d$	0.87 (8.09×10^{-2})$_d$
Epithelial to mesenchymal transition	0.85 (5.83×10^{-2})$_d$	0.87 (7.31×10^{-2})$_d$	0.88 (8.47×10^{-1})$_d$	0.94 (4.61×10^{-1})$_d$	1.00 (1.39×10^{-1})$_d$	1.00 (9.02×10^{-1})$_d$	0.96 (8.35×10^{-1})$_d$
Homophilic cell adhesion	0.90 (1.11×10^{-1})$_d$	0.85 (8.21×10^{-1})$_d$	0.95 (2.00×10^{-1})$_d$	0.94 (8.94×10^{-1})$_d$	0.97 (9.86×10^{-1})$_d$	0.98 (1.02×10^{-1})$_d$	1.00 (8.28×10^{-7})$_d$
Positive regulation of bone mineralization	0.86 (2.54×10^{-2})$_d$	0.92 (8.37×10^{-3})$_d$	0.87 (1.31×10^{-1})$_d$	1.00 (4.64×10^{-1})$_d$	0.95 (6.22×10^{-2})$_d$	0.98 (1.17×10^{-1})$_d$	0.87 (1.37×10^{-1})$_d$

a) The subscript u or d represents *P*-values for up-regulated or down-regulated genes in the modules.

human cancers. Indeed, the functional modules found here are significantly associated with these hallmarks.

(i) The module "positive regulation of I-kappaB kinase/ NF-kappaB cascade" is often regarded as pro-oncogenic signal which is a part of growth signals (the first hallmark). Inhibitors of the NF-kappaB pathway were recently used with success as treatment against cancer [32]. In this module, NF-kappaB p65, an oncogene activated in gastric carcinoma tissue [33], was measured in six datasets except the dataset for prostate cancer. We found it was up-regulated in five datasets but down-regulated in liver cancer (Table 2).

(ii) The module "transcription initiation from RNA polymerase II promoter" plays an important role in cancer cell proliferation. In this module, alternation of the function of E2F transcription factors will render cells insensitive to antigrowth factors, which is the second acquired capability of cancer cells. We found that in this module, transcription factors such as E2F2, E2F3 and many other transcription initiation factors were significantly up-regulated in tumor samples in all the seven datasets (Table 2).

(iii) The module "cell cycle arrest" is a process by which the cell cycle is halted during one of the normal phases. For example, cells in G1 phase suffering DNA damage do not enter S phase [34]. This alternation can make cancer cells to evade apoptosis, the third hallmark of cancer. In this module, p53 (Tp53) was up-regulated in six of the seven datasets except in prostate cancer.

(iv) The module "water transport" is also associated with cancer. The mechanism of cell migration, which is an initial step in angiogenesis (the fifth hallmark), can be determined by the osmotically driven water properties of the channel [35]. Saadoun *et al*. [36] showed that deletion of water channel aquaporin-1 (AQP-1) may reduce endothelial cell migration. We found the AQP-1 gene was down-regulated in all the seven datasets and the "water transport" module was significantly down-regulated in five of the seven datasets except the two colorectal cancer datasets (Table 2).

(v) Another three functional modules, "epithelial to mesenchymal transition", "regulation of bone mineralization" and "homophilic cell adhesion", are involved in the "tissue invasion and metastasis" capability (the sixth cancer hallmark). During epithelial-mesenchymal transition (EMT), epithelial cells become independent of their neighbors and can move freely. In this way, tumors derived from epithelial cells become motile and may invade lymph or blood vessels [37], accompanied by cell-cell adhesion. The module "homophilic cell adhesion" involves the tethering of cells to their surroundings in a tissue. It is known that E-cadherin, a cell-cell adhesion receptor, is an important determinant of tumor progression, serving as a suppressor of invasion and metastasis in many cancers [38]. We found the E-cadherin was up-regulated in prostate and lung cancer datasets, but down-regulated in liver, gastric and colorectal cancer datasets. "Bone mineralization" module is important in cancer metastasis [39]. In this module, bone morphogenetic protein 2 (BMP2) genes were down-regulated in all the seven datasets, which may induce medulloblastoma cell apoptosis [40].

Notably, the result that the seven modules are involved in five of the six cancer hallmarks suggests that these modules are common in tumorigenesis. The fourth hallmark (limitless replicative potential) characterized with telomere maintenance was not found by our approach largely because it was only annotated with about 20 genes. However, we found the classification accuracy of this module was beyond 90% for lung, colorectal and gastric cancer.

3 Discussion

By analyzing seven datasets for the five cancer types, we found that a wide range of functional modules were altered globally in cancers and thus had high disease classification abilities. Notably, for evaluating the disease relevance of the

functional expression of genes in a GO term, we built a classifier by using all the genes annotated to a GO term instead of selecting a few DEGs. Classifiers built in such a way are not affected by the uncertainty of DEGs selection, though they might not be able to achieve the best performance in a dataset. Actually, the classification accuracy and robustness could be improved by comparing algorithms based on different hypotheses. Especially, by integrating gene expression profiles with other data sources, we may also improve the diagnosis of many cancers. For example, by integrating gene expression profiles with protein interactions, Chuang [41] identified protein interaction subnetworks with coherent expression patterns of their component genes, which were efficient in classifying cancer metastasis. Similarly, Taylor [42] showed that the change of global modularity of human interactome can be used to assess cancer outcome. Improving algorithms by integrating diverse data sources to increase the classification accuracy and robustness for cancers deserves our future work.

Our results also suggest that alterations in some biological pathways may be specific to a cancer. For example, "vesicle" associated modules can achieve a higher classification accuracy rate for prostate caner whereas "viral genome replication" and "virion transport" modules can achieve a higher classification accuracy rate for liver cancer. However, whether this difference of modules in classification accuracy rate for different cancers could suggest some cancer specificity modules needs to be further studied. In our recent study [12,43], we found gene expressions are widely changed and highly consistent in different cancers when considering correlated gene expressions, suggesting that the expression patterns of the altered genes in different cancers tend to be similar. Because of the correlation structure of gene expressions, it would be hard to identify cancer specific expression changes in current small-scaled microarray experiments. Therefore, in this study, we only highlighted functional modules shared by various cancers.

Finally, we note that the identified modules such as cell cycle, signaling and growth regulatory modules are similar with the modules identified by Segal *et al*. using gene enrichment analysis [44]. However, our findings do not rely on individual genes whose selection is hardly reproducible. These modules can serve as functional signatures for building robust cancer classifiers. On the other hand, the traditional task of extracting biologically important individual genes still warrants investigations. For example, in our future work, we may focus on studying functional modules enriched with highly altered genes in cancer [45] to find individual genes which may play key roles in tumorigenesis.

This work was supported by the National Natural Science Foundation of China (Grant Nos. 30170515, 30370388 and 30970668).

1 Tusher V G, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA, 2001, 98: 5116–5121
2 Jeffery I B, Higgins D G, Culhane A C. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics, 2006, 7: 359
3 Chen J J, Wang S J, Tsai C A, *et al*. Selection of differentially expressed genes in microarray data analysis. Pharmacogenomics J, 2006
4 Cui X, Churchill G A. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol, 2003, 4: 210
5 Pavlidis P, Li Q, Noble W S. The effect of replication on gene expression microarray experiments. Bioinformatics, 2003, 19: 1620–1627
6 Zhang M, Yao C, Guo Z, *et al*. Apparently low reproducibility of true differential expression discoveries in microarray studies. Bioinformatics, 2008, 24: 2057–2063
7 Guo L, Lobenhofer E K, Wang C, *et al*. Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nat Biotechnol, 2006, 24: 1162–1169
8 Shi L, Reid L H, Jones W D, *et al*. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol, 2006, 24: 1151–1161
9 Allison D B, Cui X, Page G P, *et al*. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet, 2006, 7: 55–65
10 Wolfe C J, Kohane I S, Butte A J. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics, 2005, 6: 227
11 Breitling R, Armengaud P, Amtmann A, *et al*. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett, 2004, 573: 83–92
12 Yang D, Li Y, Xiao H, *et al*. Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. Bioinformatics, 2008, 24: 265–271
13 Khalil I G, Hill C. Systems biology for cancer. Curr Opin Oncol, 2005, 17: 44–48
14 Morley M, Molony C M, Weber T M, *et al*. Genetic analysis of genome-wide variation in human gene expression. Nature, 2004, 430: 743–747
15 Guo Z, Zhang T, Li X, *et al*. Towards precise classification of cancers based on robust gene functional expression profiles. BMC Bioinformatics, 2005, 6: 58
16 Barrett T, Troup D B, Wilhite S E, *et al*. NCBI GEO: mining tens of millions of expression profiles-database and tools update. Nucleic Acids Res, 2007, 35: 760–765
17 Garber M E, Troyanskaya O G, Schluens K, *et al*. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci USA, 2001, 98: 13784–13789
18 Bhattacharjee A, Richards W G, Staunton J, *et al*. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA, 2001, 98: 13790–13795
19 Galamb O, Gyorffy B, Sipos F, *et al*. Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature. Dis Markers, 2008, 25: 1–16
20 Sabates-Bellver J, Van der Flier L G, de Palo M, *et al*. Transcriptome profile of human colorectal adenomas. Mol Cancer Res, 2007, 5: 1263–1275
21 Lapointe J, Li C, Higgins J P, *et al*. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci USA, 2004, 101: 811–816
22 Chen X, Leung S Y, Yuen S T, *et al*. Variation in gene expression patterns in human gastric cancers. Mol Biol Cell, 2003, 14: 3208–3215
23 Chen X, Higgins J, Cheung S T, *et al*. Novel endothelial cell markers in hepatocellular carcinoma. Mod Pathol, 2004, 17: 1198–1210
24 Rhodes D R, Chinnaiyan A M. Integrative analysis of the cancer transcriptome. Nat Genet, 2005, 37: 31–37
25 Diehn M, Sherlock G, Binkley G, *et al*. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression

data. Nucleic Acids Res, 2003, 31: 219–223
26 Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, 25: 25–29
27 Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumours using gene expression data. J Am Statisti Associ, 2002, 97: 77–87
28 Zhang H, Yu C Y, Singer B. Cell and tumor classification using gene expression data: construction of forests. Proc Natl Acad Sci USA, 2003, 100: 4168–4172
29 Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics, 2006, 22: 1600–1607
30 Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics, 2007, 23: 257–258
31 Hanahan D, Weinberg R A. The hallmarks of cancer. Cell, 2000, 100: 57–70
32 Olivier S, Robe P, Bours V. Can NF-kappaB be a target for novel and efficient anti-cancer agents? Biochem Pharmacol, 2006, 72: 1054–1068
33 Long Y M, Ye S, Rong J, *et al*. Nuclear factor kappa B: a marker of chemotherapy for human stage IV gastric carcinoma. World J Gastroenterol, 2008, 14: 4739–4744
34 Hartwell L H, Kastan M B. Cell cycle control and cancer. Science, 1994, 266: 1821–1828
35 Rosengren S, Henson P M, Worthen G S. Migration-associated volume changes in neutrophils facilitate the migratory process *in vitro*. Am J Physiol, 1994, 267: 1623–1632
36 Saadoun S, Papadopoulos M C, Watanabe H, *et al*. Involvement of aquaporin-4 in astroglial cell migration and glial scar formation. J Cell Sci, 2005, 118: 5691–5698
37 Larue L, Bellacosa A. Epithelial-mesenchymal transition in development and cancer: role of phosphatidylinositol 3' kinase/AKT pathways. Oncogene, 2005, 24: 7443–7454
38 Jeanes A, Gottardi C J, Yap A S. Cadherins and cancer: how does cadherin dysfunction promote tumor progression? Oncogene, 2008, 27: 6920–6929
39 Hynes R O. Metastatic potential: generic predisposition of the primary tumor or rare, metastatic variants—or both? Cell, 2003, 113: 821–823
40 Hallahan A R, Pritchard J I, Chandraratna R A, *et al*. BMP-2 mediates retinoid-induced apoptosis in medulloblastoma cells through a paracrine effect. Nat Med, 2003, 9: 1033–1038
41 Chuang H Y, Lee E, Liu Y T, *et al*. Network-based classification of breast cancer metastasis. Mol Syst Biol, 2007, 3: 140
42 Taylor I W, Linding R, Warde-Farley D, *et al*. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol, 2009, 27: 199–204
43 Zhang M, Zhang L, Zou J, *et al*. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. Bioinformatics, 2009, 25: 1662–1668
44 Segal E, Friedman N, Koller D, *et al*. A module map showing conditional activity of expression modules in cancer. Nat Genet, 2004, 36: 1090–1098
45 Campagne F, Skrabanek L. Mining expressed sequence tags identifies cancer markers of clinical interest. BMC Bioinformatics, 2006, 7: 481